DOCUMENT RESUME

ED 479 597 IR 022 201

AUTHOR	Pritchard, David E.; Morote, Elsa-Sofia
TITLE	Reliable Assessment with CyberTutor, a Web-Based Homework Tutor.
PUB DATE	2002-10-00
NOTE	8p.; In: E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education. Proceedings (7th, Montreal, Quebec, Canada, October 15-19, 2002); see IR 022 093.
AVAILABLE FROM	Association for the Advancement of Computing in Education (AACE), P.O. Box 3728, Norfolk, VA 23514. Tel: 757-623-7588; e-mail: info@aace.org; Web site: http://www.aace.org/DL/.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Computer Assisted Testing; Computer Software Development; *Computer System Design; Data Collection; Educational Technology; Educational Testing; *Homework; *Individual Testing; Instructional Effectiveness; Intelligent Tutoring Systems; *Student Evaluation; Tutoring; Tutors

ABSTRACT

This paper demonstrates that an electronic tutoring program can collect data that enables a far more reliable assessment of students' skills than a standard examination. Socratic electronic homework tutor, CyberTutor can integrate effectively instruction and assessment. CyberTutor assessment has about 62 times less variance due to random test error than a three-hour final examination, and 53 times less variance than 12 weekly tests. This improvement results from the combination of three factors. First, the typical student interacts with CyberTutor 16 times longer than the three-hour final exam and 10 times more than 12 weekly tests together, diminishing the effect of lucky guesses and careless errors. Secondly, by considering requests for hints, solutions, and, importantly, the number of wrong answers and the time the student takes to complete each part of the problem, CyberTutor can make a better determination of the student's skill on a problem. Lastly, if a problem is too difficult for the student as presented, CyberTutor's hints and responses to wrong answers systematically adapt the difficulty of the problem to the student's skill level. (Contains 10 references.) (Author/AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Reliable Assessment With CyberTutor, A Web-Based Homework Tutor

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

G.H. Marks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

David E. Pritchard and Elsa-Sofia Morote
Research Laboratory of Electronics
Massachusetts Institute of Technology
United States
dpritch@mit.edu , smorote@mit.edu

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract: Socratic electronic homework tutor, CyberTutor can integrate effectively instruction and assessment. CyberTutor assessment has about 62 times less variance due to random test error than a three-hour final examination, and 53 times less variance than twelve weekly tests. This improvement results from the combination of three factors. First, the typical student interacts with CyberTutor 16 times longer than the 3-hour final exam and 10 times more than 12 weekly tests together, diminishing the effect of lucky guesses and careless errors. Secondly, by considering requests for hints, solutions, and importantly the number of wrong answers and the time the student takes to complete each part of the problem, CyberTutor can make a better determination of the student's skill on a problem. Lastly if a problem is too difficult for the student as presented, CyberTutor's hints and responses to wrong answers systematically adapt the difficulty of the problem to the student's skill level.

ED 479 597

Introduction

It is the end of semester and you have two students waiting to talk with you about their score in the final exam. You know what they want. Are you sure they deserve to fail? Are you wondering if that student who miraculously passed, deserved to pass? You are a fair teacher, and put all your effort into constructing a fair final exam but have you ever tested how reliable your final exam was? Do you really believe that a 3-hour final exam will accurately reflect the precise "skill" level of every student?

This paper demonstrates that an electronic tutoring program can collects data that enables a far more reliable assessment of students' skills than a standard examination. The following scenario illustrates the basic idea: Imagine that a rich ship-owner has hired Socrates to tutor his children. At the end of the month he desires to assess the amount they have learned. Would you advise him to:

a) Administer a standardized hour-long test to the children? or
b) Ask Socrates how much they have learned?

The answer is straightforward. Clearly Socrates' many hours of individual interaction enables him to form a far more accurate and nuanced assessment of each student than could be provided by a one-hour test. Moreover, this integrated assessment is continuously updated and is used by Socrates to adjust the pace and content of the session. Importantly, this continuous integrated assessment is available without the stress and time loss of traditional testing.

Tutoring with Assessment

The recent advances on the Internet and World Wide Web have brought attention to the world of Web-based education. There is an increasing demand for use of computerized educational aids by students and faculty (Cao & Bengu, 2000). Teachers are not only using electronic homework systems but recently there has been growing interest in using tutoring agents in education systems (Boy, 1995). The educational impact of real-time processing and reporting results of diagnostic electronic homework of results is recognized by Just in Time Teaching (Novak et al., 1999). It enables classroom instruction to be efficiently tuned to students' needs. The essential element is feedback between the web-based tutoring and classroom activities. The positive effects on learning of these web-based electronic homework tutors have been confirmed by several studies such as Mestre et

IR022201

BEST COPY AVAILABLE

al. (2000); Ogilvie (2000); Thoennessen & Harrison (1996) and Pritchard & Morote (2002a). Some of these electronic systems also eliminate the task of grading; enabling teachers to shift efforts to instruction, but only few offer an integrated system of tutoring and assessment.

Researches at Massachusetts Institute of Technology have built "CyberTutor" a Socratic electronic homework tutor. CyberTutor has a fine-grained assessment algorithm which permits the identification of skill deficiencies in foundational skills as each student moved through the course. This would allow tutors (or perhaps additional CyberTutor problems) to remedy the deficiency before it causes irreparable harm to the student. Additionally, knowing the class' skill profile on each detailed topic could allow the teacher's next lecture to review topics that were not being well learned.

The CyberTutor analog of Socrates' assessment is a skill profile – that student's skill rating on each of a set of predetermined topics. A topic could be an item in the regular syllabus like momentum, but it could also be a foundational skill like vector components, a general skill like using dimensions and units, or a conceptual topic like Newton's concept of force.

Because it is continuously updated, skill profiling has the following potentially revolutionary benefits for the student and the teacher

1) It is vastly fairer to use as the basis of a grade because of its dramatically smaller random error
2) Students who have gaps in their foundational skills can be identified and helped early in the course
3) The skill level of the class on the current topics can guide the teacher's allotment of class time – a new form of Just in Time Teaching.
4) The assessment is sufficiently detailed to be used to predict the student's grade on a standard test, for example a midterm or the AP physics test.
5) Teachers can therefore dispense with some examinations and increase class time

Skill profiling can have a revolutionary impact if it is used as a grade (e.g. your grade on the momentum homework is your skill at momentum on Thursday night at 10PM). This could replace the students' current strategy of "avoid lost points" with the much improved one "learn the most."

Methodology

Using data of Introductory Newtonian Mechanics, course 8.01 teaching at spring 2001 at MIT we evaluated reliabilities of the final exam, weekly tests, and CyberTutor. Over 90% of the students taking 8.01 in the spring term have previously attempted this course without being able to learn the problem solving skills demanded by the 8.01 final examination. Besides course elements such as class participation, written homework, group problem solving, and a Socratic electronic homework tutor, the students were given 12 weekly tests during the semester and a final exam.

We note in passing that this particular final exam had three features that should give it less error than typical physics finals. A standard test (Mechanics Baseline test) was used for 1/3 of the Final. The remainder was divided into basic skill problems and complex problems which accentuate the performance of different students. The total test requested 47 specific responses from the students. In comparison, the 8.01 final in the preceding semester required only 28, and therefore probably had larger relative statistical fluctuations.

Testing Reliability

By definition, someone taking a test or other assessment measure achieves the "observed score." The observed score is thought to be composed of two different components: The true score, an exact measure of the amount of the trait or skill assessed, offset by a random error score (Observed score = True score + Error score). The error may result from individual student responses to particular questions (Manson & Bramle, 1997) or external conditions such as fatigue and noise.

The variance of the observed score is composed of the variance of true scores and the variance of error. Then, $\sigma^2_o = \sigma^2_t + \sigma^2_\epsilon$, where, σ^2_o = variance in observed scores, σ^2_t = variance in true scores, and σ^2_ϵ = variance of error

The "reliability" of a test or assessment measures its repeatability or freedom from random testing errors. An assessment is considered reliable if it would give us the same result over and over again. By definition, the reliability is the ratio of variance of the true skill divided by the total variance of the measurement (observed variance), σ^2_t / σ^2_o. The reliability provides a direct guide to the reproducibility of the grade between two similarly

prepared tests (odd or even numbered questions) on equivalent material. If reliability = 1, there is no error in the test.

A standard method of determining the reliability of an assessment is to divide the test questions into two subtests, for example the odd and even questions if they were equal in coverage of topics and difficulty. If the resulting subtests determine skill with perfect reliability, each student would receive the same score on both subtests, and a scatter plot of each student's scores on the "odd (a)" and "even (b)" subtests would lie on the diagonal. We compute the reliability of the various assessments considered here from the *uneven split reliability formula* (see Pritchard & Morote, 2002a for derivation). This method does not assume that the halves have equal standard deviations, an assumption of the Spearman-Brown formula[1].

$$\text{Reliability} = 1 - (1+\beta)^2 / 4\beta \bullet \sigma^2_{(a-b)} / \sigma^2_{(a+b)} + (1-\beta)^2 / 4\beta \qquad (1)$$

where $\beta = \mu_a / \mu_b$, the ratios of the means of the two subsets. In this formula, the major factor that reduces the reliability below 1.0 is the mean variation of the difference between the scores on the even and odd collections of problems, $\sigma^2_{(a-b)}$.

Testing Validity

Validity relates to the meaning of the scores and the ways we use scores to make decisions (American Psychological Association, 1985). Validity addresses the issue of how well do these scores reflect the physics achievement. One way to empirically evaluate validity of the CyberTutor data is to study its "predictive validity", its ability to predict some subsequent performance. The correlation coefficient between prediction and subsequent performance now becomes the validity coefficient.

The Value of Electronic Homework Tutor as an Assessment Instrument

Reliability of CyberTutor

The reliability method was applied first to the MIT 8.01 final exam for 2001. Figure 1 (left panel) shows some lack of consistency between the performances on the two subtests in the Final exam. This reflects the error on the two subtests: they covered the same material, but random errors are considerable. Of course, adding the two halves together reduces the relative size of the error and results in reliability of 0.85 as computed from (1). In the same way, we computed the reliability of the 12 weekly tests resulting in a reliability of 0.87. The reliability of these tests, 0.85 and 0.87, is a moderate reliability in a test. Generally reliability above 0.90, typical of ETS[2] tests, is considered good for a test in education and behavioral sciences (Manson & Bramle, 1997). For example, in the case of the final exam, its reliability means that 85 percent of the variance in the observed scores is due to true score variance. The error causes 15 percent of the observed variance. This 15 percent error will offer, lead to a misdetermination of who should pass and fail as we discuss later.

If someone is highly skillful, he generally makes fewer errors and takes less time to successfully complete a task than a less skilful person. Moreover, any individual has the option to work more quickly at the expense of making more errors or vice versa. For the electronic tutor, we developed an assessment algorithm that depends on the number of right responses and hints, subparts, and number of solutions requested. We also included additional variables (time and wrong answers) as well as all other variables available from the log of each student's interactions with CyberTutor including among others, number of problems worked, etc. We call this assessment algorithm the "Enhanced CyberTutor Assessment"

The CyberTutor algorithm is computed from the even problems and the odd problems for each student (Figure 1, right panel). The curvature indicates a small excess of easier problems in the even group. The reliability from (1) of the combined assessment is 0.9976. This is a fantastic improvement over the final examination and weekly tests. That means that 99.8 percent of the observed variance in scores is due to true score variance. Overall, it has 62 times less error than final exam and 58 times less error than the all weekly tests together. It provides nearly error-free assessment.

---

[1] Spearman-Brown Reliability = 2 • [correlation (Odd, Even)] / [1 + correlation(Odd, Even)]

[2] Educational Testing Service, e.g. tests: GMAT, GRE, TOELF and PRAXIS.

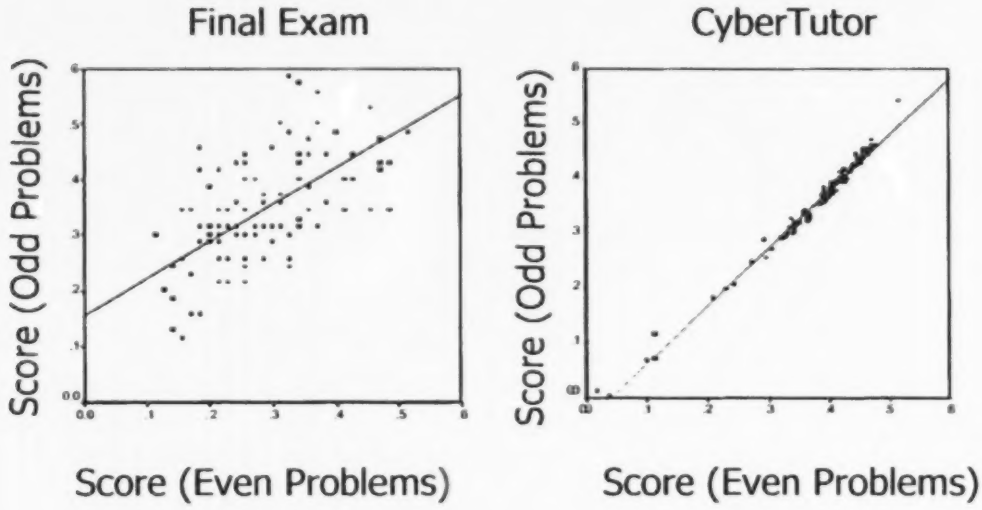

Figure 1: Grades of each individual student on the two subtests (odd -y axis - and even – x axis), comprising the CyberTutor and MIT 8.01 Final Exam in 2001.

Table 1 shows the differences among Final Exam, Weekly Tests and CyberTutor. The Enhanced CyberTutor algorithm shows higher reliability than the other tests. In addition, by eliminating marginally discriminating problems[3], we were able to increase the reliability to 99.92 percent. We called this "CyberTutor -" (the negative sign indicates that we took out all negative discriminating problems). "CyberTutor -" has 183 times less variance than the final exam. Table 1 also shows the difference of the variance per unit of assessment time The "CyberTutor -" assessment has over an order of magnitude less error variance per time than the final in this regard.

Table 1: Summary

	Reliabilities	s^2_e/s^2_t per unit time	s^2_e/s^2_t	s_e/s_t
Final exam	0.85	0.011	0.17	0.42
Weekly tests	0.873	0.014	0.144	0.380
E. CyberTutor	0.9976	0.0024	0.0024	0.049
CyberTutor -	0.9992	0.0008	0.0008	0.028

The final exam, which was carefully developed, contains so much error that 1/4 of the students assigned failing grades should has pass, and an equal number of students who should have failed are assigned a passing grade. Thus, for every two students who took the final exam one was misdiagnosed. A similar case occurs using weekly tests.

To see how this arises, look at Figure 2, where failure is set at one standard deviation below average, which fails 16 percent of the students, the typical failure rate for 8.01 course. The dashed line shows the true skill distribution inferred from the measured error and the observed skill distribution (heavy line), which is slightly broader. The number of students failed versus their true skill level, is indicated by the small curve centered on skill = 40. (The width of this curve is very nearly the width of the error.)

[3] By performing item analysis using item discrimination. It measures how well an item distinguishes between the students who understand the content universe if the test well and those who do not.

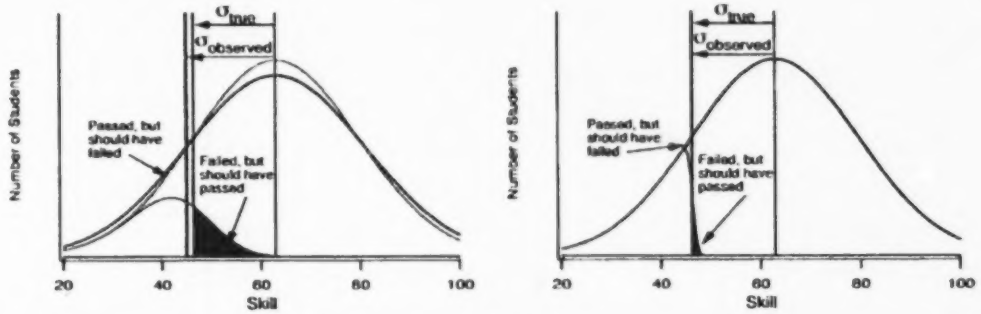

Figure 2: Fairness in the Final Exam versus CyberTutor

Below skill = 35 the failed students and the true skill distribution were coincident, indicating that all students whose skill is this low fail. Failed student curve drops well below the true skill distribution between 35 and 46 (the selected true skill for passing) indicating that some students (lighter shaded area) whose ability is below passing will pass. The distribution of failed students does not drop immediately to zero above true skill = 46, these (darker shaded area) whose true ability is passing, will fail the test.

If the much more accurate Enhanced CyberTutor Assessment were used to set pass-fail instead of the final exam, only 1 percent of the students failed would be misdiagnosed (Figure 2, right panel). In a class of 100, only one student would be misdiagnosed, and his true skill would almost certainly lies only 1 or 2 points away from the pass-fail line.

Validity of CyberTutor

The skill profile can be used to predict each student's grade on a standard test. As a first step in this direction, we developed an algorithm to predict students' grades on the "adjusted" MIT final exam. The adjustment reflected the findings from item analysis that both exam questions concerning angular dynamics were poor measures of overall skill. It is well known that many students "hit the wall" of misunderstanding here, and while we believe that an examination of CyberTutor data concerning angular dynamics would improve our predictive ability on these problems, it was cleaner to eliminate them for this study of validity.

The final prediction (Figure 3) yields a value for r, the predictive validity coefficient, of 0.71, which is highly statistically significant[4]. In fact this correlation coefficient would be higher were in not for the 15% unreliability of the final exam. This suggests that the "real" validity coefficient is around 0.77. This level suggests that a variant of the "Enhanced CyberTutor algorithm" would enable us to predict the students' skills with less error than a one-hour test such as a midterm could determine it.

We are optimistic that a prediction based on a detailed skill profile instead of the overall performance could do considerably better at predicting the final exam scores and conclude that CyberTutor as a measurement instrument is not only extremely reliable but also has significant validity. Indeed there is every reason to believe that it would give a better indication of students in trouble at midterm time than the midterm itself.

In this section, we have begged the question of the validity of adopting the final exam as a criterion of validity (e.g. only about 1/4 of the final exam involved conceptual questions). In the future we hope to use the performance in the subsequent required physics course as another measure of validity.

[4] The minimum value of r required to be statistically significant at the 0.05 levels a sample size between 100 and 1000 is 0.197 (Hopkins and Glass, 1978 p. 409).

BEST COPY AVAILABLE

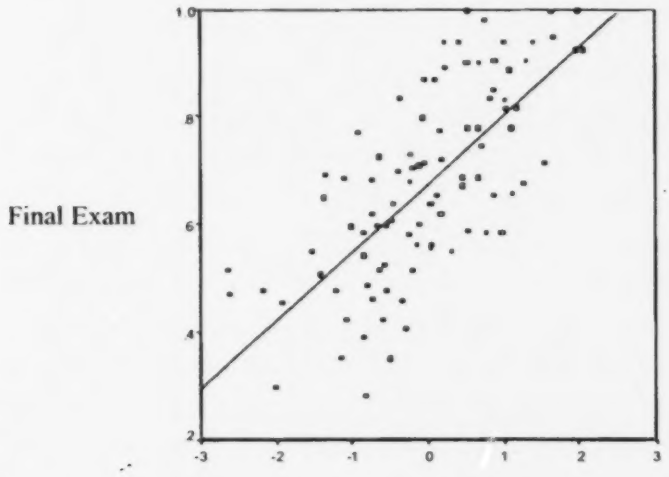

Figure 3: Prediction of scores on Adjusted Final Exam from CyberTutor data, r = 0.71. The horizontal axis is deviation from CyberTutor prediction in standard deviations

Discussion

We have shown that a web-based electronic homework tutor collects data that enables an assessment and tutoring and gives us a better way to deal fairly with students' capabilities. CyberTutor provides objective measurements on which educational decisions can be based. It can provide means for improving and enhancing feedback, motivation, and over learning.

The dramatic reduction in the variance of the error results from combination of three factors. First, the typical student interacts with CyberTutor for at least 48 hours per term, 16 times longer than the 3-hour final exam and 10 times more than 12 weekly tests together. This reduces the error variance due to the general statistical rule that the variance of error approximate 1/T, where T is the total time of the measurement.

Secondly, by considering requests for hints, solutions, and importantly the number of wrong answers and the time the student takes to complete each part of the problem, CyberTutor can make a more reproducible determination of the student's skill on a problem. Lastly if a problem is too difficult for the student as presented, CyberTutor's hints and spontaneous responses to wrong answers systematically adapt the difficulty of the problem to the student's skill level, and the CyberTutor assessment algorithm accurately determines that level. In contrast, too difficult written problems provide only an upper bound on the student's skills.

Tutor-based assessment has such low error that it is capable of providing accurate assessments on a number of pre-selected topics providing the detailed skill profile. If we took 20 topics – say the major topic of the 13 weeks of the course plus 7 foundational topics; the assessment on each topic would be twice times as reliable as the overall assessment of the current final. Even if we increased the skill resolution to include each of the major weekly topics (i.e. perhaps 180 topics) the variance on each would be small enough to identify those students requiring immediate assistance on a particular topic. By basing the assessment on a more careful analysis of wrong answers (many wrong answer reflect omission of a particular step or reflect use of a particular wrong assumption rather than just a careless error) we project that we can improve the CyberTutor assessment sufficiently to measure skill across this fine grid with the accuracy that the current final determines the overall skill of the students.

References

Boy, G. (1995). *Software agents for cooperative learning*, in J.M. Bradshaw (Ed.) AAAI Press/The MIT Press.

Cao, L., & Golden, B. (2000). *Web-based agents for reengineering engineering education*. J. Educational Computing Research, 23, 4, 421-430.

Ebel, R., & Frisbie, D. (1991). *Essentials of educational measurement*. New Jersey: Prentice Hall.

Manson, E., & Bramle, W. (1997). *Research in education*. Sydney: Brow & Benchmark Publishers.

Mestre, J., Dufrense, R., Hart, D., & Rath, K. (2000). The effect of web-based homework on test performance in large enrollment introductory physics courses. *Journal of Computers in Mathematics and Science Teaching*

Novak, G., Patterson, E., & Gavrin, A. (1999) *Just-in-Time Teaching : Blending active learning with web technology*. Prentice Hall Series in Educational Innovation.

Ogilvie, C. (2000). Effectiveness of different course components in driving gains in conceptual understanding. Cambridge, Internal report, Department of Physics at MIT. [on-line] URL: http://torrseal.mit.edu/effedtech/

Pritchard, D. & Morote, E. (2002a) *99.8% Reliable Assessment with CyberTutor. Physics Department, Report to Massachusetts Institute of Technology*. [on-line] URL: http://torrseal.mit.edu/effedtech/

Pritchard, D., & Morote, E. (2002b). *What course elements correlate with improvement on tests in introductory Newtonian mechanics?* NARST Annual Meeting New Orleans, Louisiana, April 7-10.

Thoennessen, M., & Harrison, M. (1996). Computer-assisted assignments in a large physics class. *Computers Educ.*, 27, 2 , 141-147.

Acknowledges

This work was supported by NSF grant PHY-9988732. We are grateful to Dr. David Kokorowski for his help on the Figures.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").